Syst. Biol. 61(4):653–660, 2012

DOI:10.1093/sysbio/sys026
Advance Access publication on February 14, 2012

Turning the Crown Upside Down: Gene Tree Parsimony Roots the Eukaryotic Tree of Life

LAURA A. KATZ[1,2,*], JESSICA R. GRANT[1], LAURA WEGENER PARFREY[1,2,3], AND J. GORDON BURLEIGH[4]

[1]*Department of Biological Sciences, Smith College, 44 College Lane, Northampton, MA 01063, USA;* [2]*Program in Organismic and Evolutionary Biology, University of Massachusetts, 611 North Pleasant Street, Amherst, MA 01003, USA;*
[3]*Present address: Department of Chemistry and Biochemistry, University of Colorado, 215 UCB, Boulder, CO 80309, USA; and*
[4]*Department of Biology, University of Florida, PO Box 118526, Gainesville, FL 32611, USA;*
[*]*Correspondence to be sent to: Department of Biological Sciences, Smith College, Northampton, MA 01063, USA; E-mail: lkatz@smith.edu.*

Received 7 June 2011; reviews returned 9 September 2011; accepted 17 January 2012
Associate Editor: Michael Charleston

Abstract.—The first analyses of gene sequence data indicated that the eukaryotic tree of life consisted of a long stem of microbial groups "topped" by a crown-containing plants, animals, and fungi and their microbial relatives. Although more recent multigene concatenated analyses have refined the relationships among the many branches of eukaryotes, the root of the eukaryotic tree of life has remained elusive. Inferring the root of extant eukaryotes is challenging because of the age of the group (∼1.7–2.1 billion years old), tremendous heterogeneity in rates of evolution among lineages, and lack of obvious outgroups for many genes. Here, we reconstruct a rooted phylogeny of extant eukaryotes based on minimizing the number of duplications and losses among a collection of gene trees. This approach does not require outgroup sequences or assumptions of orthology among sequences. We also explore the impact of taxon and gene sampling and assess support for alternative hypotheses for the root. Using 20 gene trees from 84 diverse eukaryotic lineages, this approach recovers robust eukaryotic clades and reveals evidence for a eukaryotic root that lies between the Opisthokonta (animals, fungi and their microbial relatives) and all remaining eukaryotes. [Eukaryotes; gene tree; molecular systematics; species tree reconciliation; tree of life.]

Early molecular phylogenetic analyses depicted the eukaryotic tree of life as ladder of unicellular lineages leading to a crown that included plants, animals, and fungi (Sogin et al. 1989; Van de Peer and De Wachter 1997; Brown and Doolittle 1999). More recent analyses of concatenated multigene sequence alignments reveal that microbial and macroscopic lineages are intermingled (e.g., Baldauf et al. 2000; Yoon et al. 2008; Burki et al. 2009; Hampl et al. 2009; Parfrey et al. 2010), but the root of the eukaryotic tree of life remains unknown. These concatenated analyses exclude paralogs, eliminating the phylogenetic information of gene duplication and loss events. Such data are important as the root of a species tree can be inferred from individual gene trees by identifying the rooting that implies the fewest duplications or duplications and losses (Gogarten et al. 1989; Iwabe et al. 1989). Gene tree parsimony (GTP) extends this concept to large collections of gene trees, seeking the rooted species tree that implies the fewest gene duplication and loss events across all gene trees (e.g., Goodman et al. 1979; Guigo et al. 1996; Maddison 1997). GTP has several advantages over phylogenetic inference methods that use alignments of concatenated putatively orthologous genes. For example, the need to identify orthologs is eliminated as all paralogs are retained and used in the inference. Furthermore, GTP approaches yield rooted topologies, and the support for alternative roots can be assessed easily. Algorithmic advances now enable effective heuristics to estimate species trees from large data sets using GTP (e.g., Wehe et al. 2008; Bansal et al. 2010; Burleigh et al. 2011).

Recent reconstructions of the eukaryotic tree of life are largely based on concatenated multigene alignments of orthologous loci (e.g., Baldauf et al. 2000; Yoon et al. 2008; Burki et al. 2009; Hampl et al. 2009; Parfrey et al. 2010). Although these approaches provide many insights into relationships among eukaryotic lineages, such analyses exclude potentially phylogenetically informative processes such as gene duplications and require accurate identification of orthologous sequences. Identifying orthologs is difficult because rampant gene duplication and differential loss create complex patterns of paralogy (Maddison 1997), which is amplified at phylogenetic levels as deep as all eukaryotes (Roger and Hug 2006). Rooting phylogenies from analyses of concatenated alignments is usually done using outgroup sequences. However, this approach is problematic for recovering root of eukaryotes because many eukaryotic genes lack clear homologs in bacteria and archaea (e.g., Mans et al. 2004; Tekle et al. 2009), and when outgroup sequences are available, the vast evolutionary distances involved can lead to systematic error associated with long branches (Felsenstein 1978).

Here, we use GTP to reconstruct eukaryotic phylogeny and identify the rooted tree of eukaryotes. We analyzed 20 genes, including all available paralogs, from up to 84 taxa (Table 1). In an earlier study, these data were pruned to retain only putative orthologs, concatenated, and analyzed using standard phylogenetic methods (Parfrey et al. 2010). Exclusion of paralogs, which is required for concatenation, eliminates potential phylogenetic information. This is because closely related species will share a common history of gene gains and losses.

We explored the effect of taxonomic sampling by using input gene trees that included sequences from

TABLE 1. Summary of gene sampling for the "15 no out" and "15 + out" data sets

	Genes																			
Taxon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Acanthamoeba castellanii	1	1	1	1	1	1	1	1	1	1	0	1	1	2	1	1	1	0	0	0
Alexandrium tamarense	1	2	2	1	1	0	1	1	3	1	0	1	1	2	0	1	1	2	0	0
Aplysia californica	1	1	1	2	3	1	1	0	1	1	0	1	1	1	1	1	1	1	1	0
Arabidopsis thaliana	1	12	1	6	2	8	1	1	3	2	1	8	6	1	1	2	2	2	2	2
Aureococcus anophagefferens	1	2	1	1	1	5	2	1	3	1	1	6	4	2	1	1	0	1	1	0
Branchiostoma floridae	1	1	1	1	2	4	2	2	1	1	1	3	1	2	1	1	1	1	1	2
Caenorhabditis elegans	1	1	1	1	8	4	1	1	1	1	1	5	2	4	1	2	1	1	1	1
Candida albicans	1	1	1	1	1	1	1	1	1	1	1	3	1	1	1	1	1	1	1	1
Capsaspora owczarzaki	1	1	1	1	1	0	1	1	1	1	0	1	2	1	0	1	1	1	0	0
Chlamydomonas reinhardtii	1	1	1	1	1	1	1	1	1	1	1	3	3	1	1	1	1	1	0	1
Ciona intestinalis	1	1	1	1	3	1	1	1	1	1	1	3	1	1	1	1	1	1	1	0
Cryptosporidium parvum	1	1	1	1	1	1	1	1	1	1	1	2	2	1	1	2	1	1	2	2
Cyanophora paradoxa	1	7	2	3	4	4	1	4	5	2	1	3	2	2	1	2	2	1	0	0
Dictyostelium discoideum	1	1	1	3	1	1	1	2	2	1	1	4	3	1	1	2	1	1	1	1
Diplonema papillatum	1	1	3	1	1	1	1	1	0	1	0	3	3	1	0	2	2	2	0	0
Drosophila melanogaster	1	4	2	1	2	4	4	1	1	1	2	6	3	2	1	1	1	1	1	1
Emiliania huxleyi	1	2	1	1	1	1	1	2	4	1	0	1	3	1	0	1	1	1	0	1
Entamoeba histolytica	1	2	1	1	1	2	2	1	1	1	2	5	2	2	1	1	1	1	2	1
Euglena gracilis	1	3	0	1	1	2	1	2	2	1	1	1	1	0	1	1	1	1	0	0
Gallus gallus	1	6	1	1	4	4	3	1	3	1	1	3	3	1	1	1	1	2	1	1
Ginkgo biloba	1	2	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0
Homo sapiens	1	5	1	5	4	9	3	1	4	4	2	5	3	2	1	2	3	2	1	1
Isochrysis galbana	1	2	1	7	2	1	1	1	2	1	1	2	2	2	1	1	1	1	0	2
Leishmania major	1	2	1	1	1	2	1	1	1	1	1	2	3	1	1	1	1	1	1	2
Malawimonas jakobiformis	1	1	1	1	1	2	1	1	0	2	0	1	1	1	1	1	1	1	0	0
Mastigamoeba balamuthi	1	1	1	1	1	0	1	1	1	1	0	1	1	1	0	1	1	1	0	1
Mesostigma viride	1	2	1	1	2	1	1	2	1	1	0	1	1	1	0	1	1	0	0	0
Micromonas pusilla	1	1	1	1	1	1	1	1	1	1	0	3	1	1	0	1	1	1	0	0
Monosiga brevicollis	1	2	1	1	1	1	1	1	1	1	1	2	3	1	1	1	1	1	1	1
Naegleria gruberi	1	1	1	1	2	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0
Nematostella vectensis	1	1	1	1	11	3	1	1	1	1	1	4	5	2	1	1	1	1	1	1
Oryza sativa	1	6	2	7	2	5	1	3	3	3	1	8	5	2	1	2	2	2	1	1
Paramecium tetraurelia	1	1	1	1	1	1	1	1	1	1	1	11	1	2	1	1	2	4	2	1
Pavlova lutheri	1	1	1	2	2	1	2	2	5	1	1	3	3	0	0	1	0	4	0	0
Perkinsus marinus	1	5	1	2	2	1	1	1	1	1	1	2	1	1	0	1	1	1	1	0
Phaeodactylum tricornutum	1	1	1	3	1	2	1	1	1	1	1	2	1	1	1	1	1	1	1	1
Phanerochaete chrysosporium	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
Physarum polycephalum	1	1	1	1	3	2	1	1	1	1	1	0	1	1	1	1	1	1	0	1
Physcomitrella patens	1	8	1	1	2	5	4	1	4	1	1	6	8	1	1	1	1	1	1	1
Phytophthora infestans	1	1	1	2	1	2	1	1	2	1	0	1	2	1	1	1	1	0	1	0
Plasmodium berghei	1	1	1	1	2	1	1	1	1	1	1	2	3	1	1	2	1	1	1	2
Porphyra capensis	1	1	1	3	2	1	1	1	1	1	0	1	0	1	1	1	1	1	1	0
Reclinomonas americana	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	2	1	1	0	0
Saccharomyces cerevisiae	1	2	1	1	2	1	1	2	3	1	1	5	1	2	1	1	1	1	1	2
Schistosoma mansoni	1	4	1	1	1	0	1	0	1	1	1	1	2	1	1	1	1	1	0	0
Schizosaccharomyces pombe	1	2	1	1	5	1	1	1	2	1	1	4	1	1	1	1	1	1	2	2
Seculamonas ecuadoriensis	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0
Sphaeroforma arctica	1	2	1	1	1	0	1	1	1	1	0	1	1	1	0	1	1	1	0	0
Strongylocentrotus purpuratus	1	2	1	2	17	8	1	1	0	1	2	7	3	1	1	1	1	1	1	1
Tetrahymena thermophila	1	1	1	1	1	1	1	1	2	1	1	6	3	1	1	1	1	2	3	1
Thalassiosira pseudonana	1	1	1	3	1	2	2	1	2	1	1	5	3	2	0	1	1	2	1	2
Theileria parva	1	1	1	1	4	1	2	1	1	1	1	2	4	1	1	2	1	1	1	1
Toxoplasma gondii	1	1	2	1	1	2	1	1	2	1	1	2	2	1	1	1	1	1	1	1
Trichomonas vaginalis	1	4	1	1	1	1	1	2	6	2	1	7	4	3	1	1	1	1	2	1
Trimastix pyriformis	1	2	1	1	2	2	1	1	1	1	0	2	2	1	0	1	1	1	0	0
Trypanosoma brucei	1	2	1	1	1	1	1	1	1	1	1	2	2	1	1	1	1	1	2	1
Ustilago maydis	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
Volvox carteri	1	1	1	1	1	2	1	1	1	1	0	1	2	1	1	1	1	1	0	1
Welwitschia mirabilis	1	4	1	1	1	1	1	1	0	1	1	1	2	1	0	1	0	1	0	0
Number of genes	59	130	65	95	128	116	75	69	97	68	45	172	128	76	44	70	62	70	44	44
Number of taxa	59	58	58	59	59	54	59	56	55	59	41	58	58	57	44	59	55	56	35	35

Notes: Gene names: 1 = small subunit rDNA, 2 = 14-3-3, 3 = 40S ribosomal protein, 4 = actin, 5 = α-tubulin, 6 = β-tubulin, 7 = elongation factor 1α, 8 = elongation factor 2, 9 = enolase, 10 = 60S ribosomal protein L9, 11 = γ-tubulin, 12 = heat shock protein 70, 13 = heat shock protein 90, 14 = S-adenosylmethionine synthetase, 15 = Rad51, 16 = 40S ribosomal protein 23, 17 = 40S ribosomal protein S15a, 18 = transport protein Sec61 subunit α, 19 = transcription factor II H, 20 = U5 snRNP/Prp8.

species represented in at least 10 or 15 of the 20 total gene alignments. GTP requires rooted gene trees, but because it is difficult to know the root of individual gene trees, we explored alternative methods to root gene trees: individual gene trees either were rooted a priori with outgroup sequences (i.e., ancient paralogs or bacterial/archaeal homologs) or were rooted by searching for the root position that minimized the implied number of gene duplications and losses during the species tree search (e.g., Gorecki and Tiuryn 2007; Sanderson and McMahon 2007; Wehe et al. 2008). We assessed the impact of these two approaches by either using: (i) all unrooted input gene trees ("no out") or (ii) a combination of 10 unrooted and 10 rooted input gene trees ("+ out").

Materials and Methods

Sequence Assembly

The data for this study expanded on the sampling from the supermatrix study of Parfrey et al. (2010) by using all available sequences from the gene families (Table 1). We aimed to sample eukaryotic diversity by including as many lineages defined by ultrastructural identities as possible, using the classification systems of Patterson (1999) and Adl et al. (2005) as guides. We excluded taxa with elevated rates of molecular evolution (e.g., *Encephalitozoon*, *Giardia*) as in Yoon et al. (2008); inclusion of these rogue taxa decreased stability of gene tree topologies and hence our ability to make inferences.

Small subunit (SSU)-rDNA sequences were hand-curated for target taxa by removing group I introns, spacers, unalignable regions, nonnuclear rDNAs, and misannotated sequences. SSU-rDNA sequences were aligned by HMMER (Eddy 2001), version 2.1.4 with default settings, taking secondary structure into account (see Parfrey et al. 2010). We note that the inclusion of SSU-rDNA had little effect on the analyses; removing this gene from our GTP analyses did not change the location of the root, although support values varied to some extent. The assembly of protein coding genes was accomplished using a custom-built pipeline of Perl and Python scripts also described in Parfrey et al. (2010). In total, our final data set included alignments of 20 loci (SSU-rDNA and 19 protein-coding genes).

Previous work indicated that GTP analyses of data sets with dense sampling (minimizing the missing data or taxa within genes) provided the strongest support (e.g., Burleigh et al. 2011). Therefore, we pruned the single-locus alignments to include (i) only sequences from taxa that were represented in at least 10 of the 20 single-locus alignments ("10" data sets) and (ii) only sequences from taxa that were represented in at least 15 of the 20 single-locus alignments ("15" data sets). Ten of the 20 loci had outgroup sequences available, and we included these outgroup sequences in the alignments for these loci. The outgroups were either ancient paralogs that predated the divergence of all extant eukaryotes or were from bacteria or archaea. The alignments that included outgroups were used to create "+ out" gene tree data sets, and the alignments that excluded outgroups were used to create "no out" gene tree data sets. Thus, in total, we created four sets of 20 single-locus alignments: "15 + out", "10 + out", "15 no out", and "10 no out". All alignments are available from the Dryad data depository (http://datadryad.org/) at doi:10.5061/dryad.11vq033v, and trees can be found in TreeBase (#12157).

Gene Tree Construction

We performed maximum likelihood (ML) phylogenetic analyses on each of the single-locus alignments using RAxML-VI-HPC version 7.0.4 (Stamatakis 2006). For all loci except SSU-rDNA, ML analyses used the Jones–Taylor–Thornton (JTT) amino acid substitution model (Jones et al. 1992) with the default settings for the optimization of individual per-site substitution rates and classification of these rates into rate categories ("JTTCAT model"). For the SSU-rDNA alignments, ML analyses used the GTRCAT nucleotide substitution model. We also performed 100 nonparametric bootstrap replicates (Felsenstein 1985) for each locus alignment, with each replicate using the same tree search as was used on the original data set. If the resulting tree included outgroup sequences, we rooted all the resulting trees using the outgroups, and then we pruned the outgroups from the trees prior to the GTP analyses. The gene tree data sets are available from the Dryad data depository (http://datadryad.org/).

GTP Analysis

Species trees were inferred from GTP analyses based on the duplication and loss model, which, given a collection of rooted gene trees, seeks the rooted species tree that implies the fewest duplications and losses across all gene trees. To estimate the optimal species tree, we used a tree search heuristic based on the rooted subtree pruning and regrafting (SPR) local search algorithm (Bansal et al. 2010), now implemented in iGTP (Chaudhary et al. 2010).

GTP evaluates species trees from rooted gene trees. However, the ML gene tree searches output unrooted gene trees, and it is often difficult to determine the root of a gene tree without obvious outgroups. One strategy for rooting gene trees for GTP analyses is to find a root that minimizes the duplication and loss cost. To do this, we gave the starting unrooted gene trees an arbitrary root. Then, for each candidate species tree, we evaluated all possible rootings for each of these gene trees and used a root that minimizes the duplication and loss cost (Gorecki and Tiuryn 2007; Sanderson and McMahon 2007). For the gene trees that were rooted with outgroups (10 trees in the "+ out" data sets), we fixed the root (i.e., constrained to bacterial sequences or ancient paralogs) and did not consider alternate roots during the tree search.

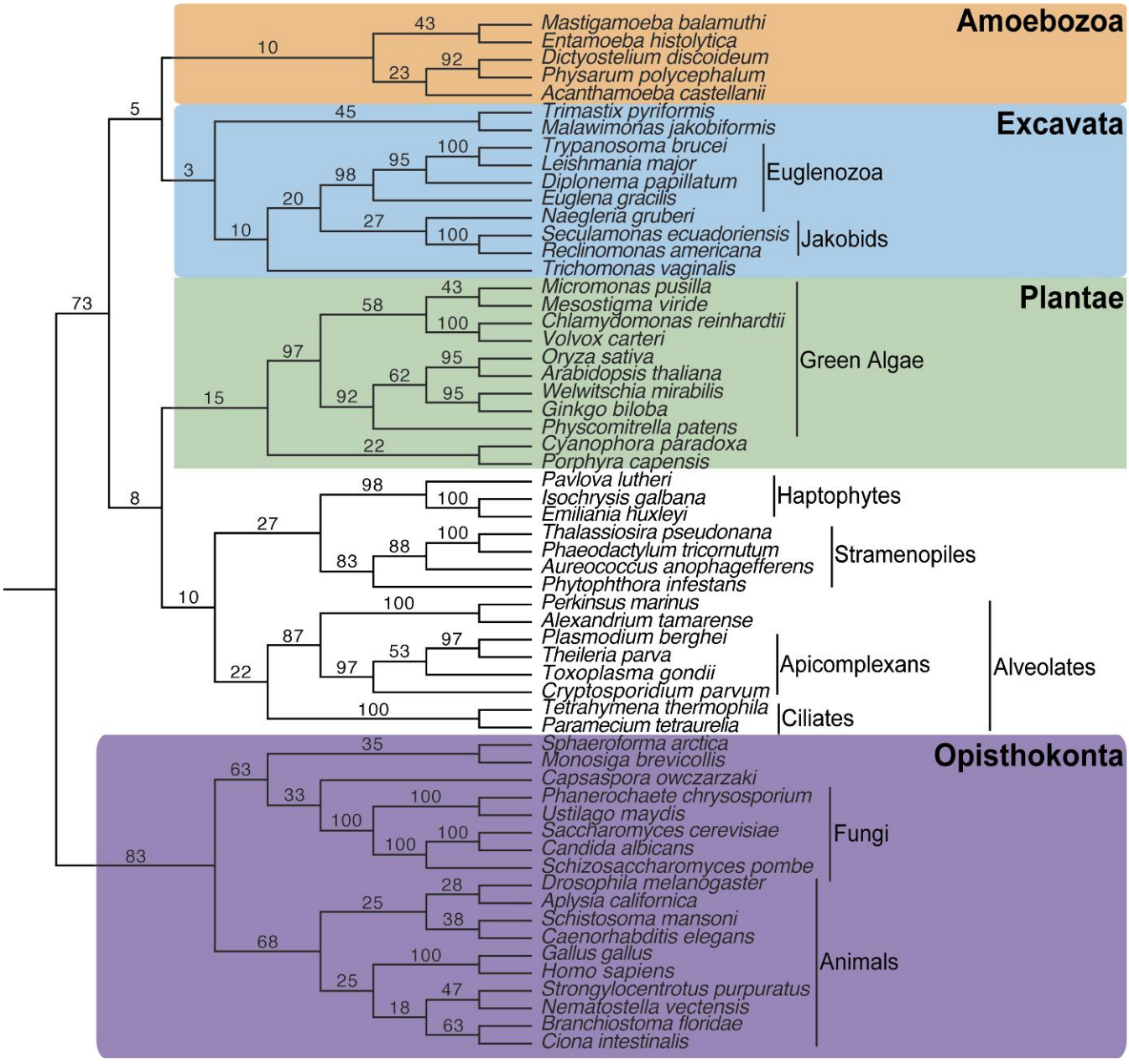

FIGURE 1. Reconciled tree of eukaryotes reveals a root between Opisthokonta and all remaining eukaryotes. Major clades indicated in colored boxes and nested clades with vertical lines, as described in Table 2. Tree estimated from analysis of 59 taxa in the "15+ out" analysis: taxa found in at least 15 of the 20 genes (Table 1) and including rooted gene trees for the 10 genes with outgroup sequences.

For each gene data set ("15 + out," "10 + out," "15 no out," and "10 no out"), we first ran 50 random stepwise addition replicates to build starting species trees and performed a local SPR tree search from each starting tree. We observed that the starting tree affected the resulting species tree estimates, and this was especially apparent with the "no out" data sets. We hypothesized that the tree searches using the "no out" data sets may be adversely affected by the arbitrary starting gene tree rootings. To address this potential problem, we rooted the genes from the "no out" data sets with a root that implies the fewest duplications and losses given the species tree estimate from the "+ out" analysis. Then, we performed 50 more SPR tree searches from random stepwise addition starting trees on the "no out" data sets that had the new starting gene tree roots (Note that we still evaluated all alternate gene tree rootings during these analyses. The only difference was the initial rooting of the gene trees.). Finally, for each data set, we examined the results of all tree search replicates to find the species trees that implied the fewest duplications and losses.

One concern about the performance of GTP, or any supertree method, is the quality of the input gene trees. Specifically, GTP analyses that use a single topology per gene may fail to incorporate tree support or account for

underlying phylogenetic signals that are not observed in a single gene tree (Page 2002). We used a variation of a "supertree bootstrapping" strategy to incorporate both uncertainty in the gene tree topology and underlying phylogenetic signals from the gene trees into the GTP analysis (e.g., Cotton and Page 2002; Burleigh et al. 2006; Joly and Bruneau 2009; Burleigh et al. 2011). In our supertree bootstrapping approach, each bootstrap replicate consisted of a GTP analysis using a single randomly selected bootstrap tree from each gene. We performed 100 replicates of the supertree bootstrap method for each data set ("15 + out," "10 + out," "15 no out," and "10 no out"). For all bootstrap replicates, we followed the same tree search protocol as was used the original data set. We summarized the results of the GTP bootstrap replicates with a greedy majority rule consensus tree.

We also examined several alternate phylogenetic hypotheses regarding the root of eukaryotes using a hypothesis-testing approach analogous to the "Templeton test" used in maximum parsimony phylogenetic analyses (Templeton 1983; Burleigh et al. 2011). For each of two rooted species trees, one designated as the null hypothesis and one as the alternate hypothesis, we calculated the gene duplication and loss score for each gene tree. For each hypothesis, each gene tree was given a rooting that minimized the duplication and loss score. We then compared the differences in the reconciliation costs for the two topologies using the Wilcoxon matched-pairs signed-ranks test. To generate alternative topologies, we constrained GTP tree searches to estimate the optimal topologies consistent with the Opisthokonta, Fungi, Unikonta, Plantae, Euglenozoa, Amoebozoa, Excavata, and SAR roots for the eukaryotic tree in accordance with hypothesized positions of the root and the major clades of eukayotes. The constrained tree searches used the same tree search protocol as we described for the original data set, with the addition of a topological constraint.

Results and Discussion

GTP Analyses of the Eukaryotic Tree of Life

The eukaryotic tree of life produced by our GTP approach recovers robust clades inferred from ultrastructural identities and previous concatenated analyses, including animals, fungi, green algae, alveolates, haptophytes, and stramenopiles (Fig. 1 and Table 2; Patterson 1999; Baldauf et al. 2000; Parfrey et al. 2006, 2010; Yoon et al. 2008; Burki et al. 2009; Hampl et al. 2009). Since GTP approaches are based on patterns of variation among gene trees, a much larger sample of gene trees will be needed to yield a strongly supported species tree. Still, with our relatively small sample of genes (20) and taxa (up to 84), the GTP analyses are generally concordant with previous studies in recovering several deeper clades (e.g., Amoebozoa, Plantae, Excavata, Opisthokonta; Fig. 1 and Table 2). This is particularly surprising given that individual gene trees differ tremendously from one another and do not support major clades of eukaryotes, as exemplified by the topology for HSP90 (Fig. 2). The congruence between the topology of the eukaryotic tree recovered in this GTP analysis and concatenated analyses fosters greater confidence in the robustness of this approach in general and in the estimate of the root that it provides.

TABLE 2. Support values for root, major clades, and nested clades from reconciled tree

	15 + out	10 + out	15 no out	10 no out
Root				
Opisthokonta, others	73	45	nm	26
Fungi, others	nm	nm	61	nm
Major clades				
Opisthokonta	83	59	nm	34
Amoebozoa	10	6	nm	4
Excavata	3	nm	nm	4
Rhizaria	na	20	na	13
SAR	na	2	na	nm
Plantae	15	nm	7	nm
Exemplar nested clades				
Animals	68	52	74	44
Apicomplexa	97	81	98	69
Ciliates	100	60	100	60
Euglenozoa	98	87	93	93
Fungi	100	76	59	72
Green algae	97	87	91	77
Haptophyta	98	89	96	89
Jakobids	100	100	100	99
Stramenopiles	83	46	84	41

Notes: Columns labeled by analyses: 15 (or 10) refers to data sets that included taxa in at least 15 (or 10) of the 20 gene trees. Outgroups (bacterial/archaeal or ancient eukaryotic paralogs) are included for 10 genes in "+ out" analyses and excluded in "no out" analyses. There are 84 eukaryotes in the "10" analyses and 59 taxa in the "15" analyses. nm = not monophyletic; na = not applicable as less than two taxa included.

The Root of the Eukaryotic Tree of Life is Between Opisthokonta and All Remaining Eukaryotes

Our reconstructions of the eukaryotic tree of life created with GTP place the root at the base of or within the Opisthokonta, a clade containing animals, fungi, and their microbial relatives. Three of the 4 analyses support a root between the opisthokonts and all remaining eukaryotes ("15 + out", "10 + out", "10 no out"; Fig. 1 and Table 2), whereas in the fourth analyses ("15 no out"), the root is placed within the opisthokonts such that Fungi are sister to all remaining eukaryotes. Supertree bootstrap support, which represents uncertainty in the gene tree topologies (e.g., Cotton and Page 2002; Burleigh et al. 2006), shows highest support for the root in the analyses limited to species in at least 15 of the gene trees with 10 of the gene trees rooted by outgroups ("15+out", 73% bootstrap support; Fig. 1 and Table 2).

Hypotheses on the root of the eukaryotic tree of life have focused on events deemed rare or are based on putatively primitive features. Rooting the eukaryotic tree of life between the Opisthokonta and all remaining eukaryotes is consistent with both initial analyses of the presence/absence of a gene fusion between

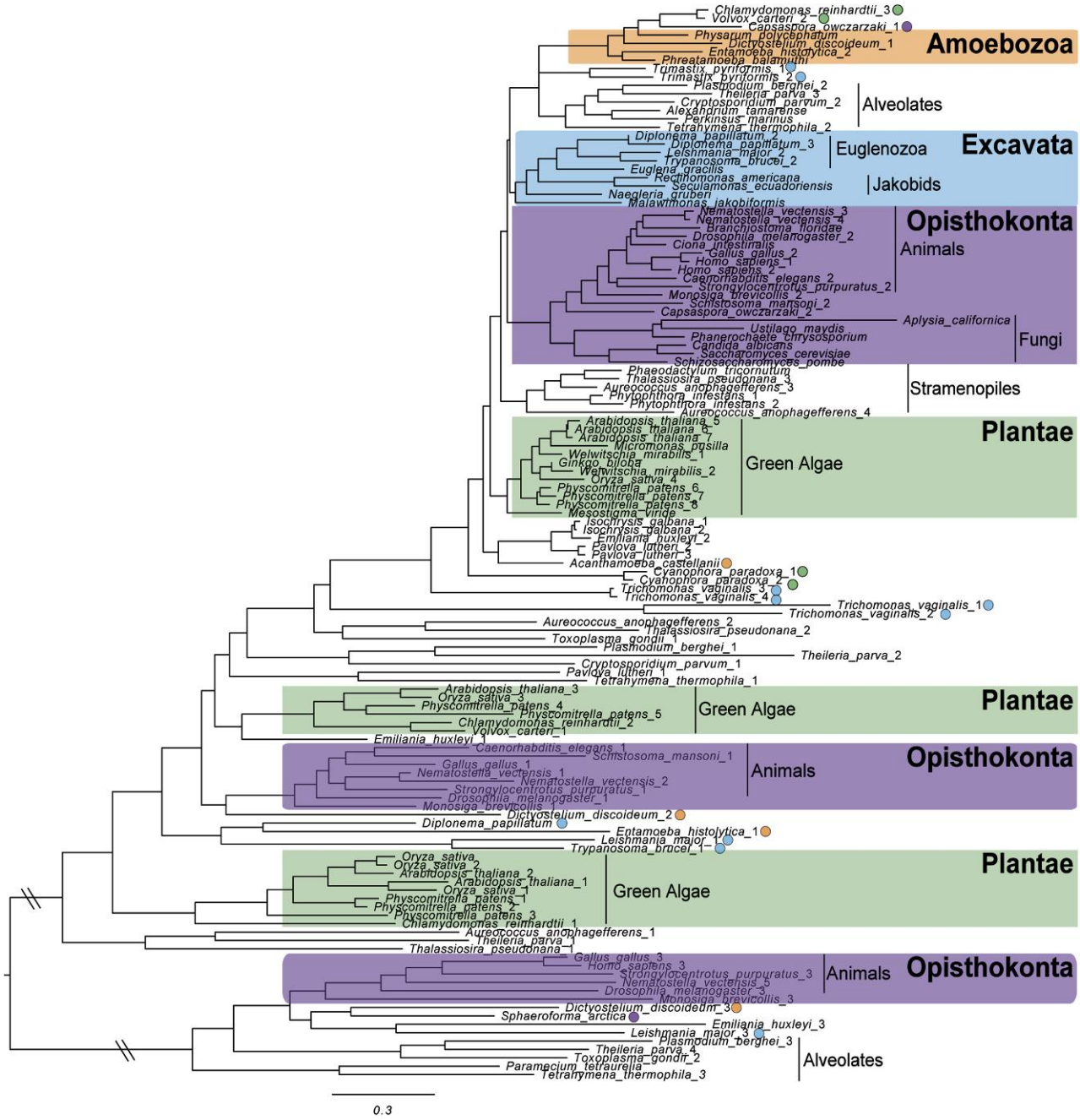

FIGURE 2. Exemplar single gene tree (HSP90) reveals a topology inconsistent with ultrastructure, previous analyses of multigene data, and with the reconciled tree (Fig.1). Major clades are not monophyletic but are shaded as in Figure 1 to highlight complexity of single gene trees.

dihydrofolate reductase (DHFR) and thymidylate synthase (TS) genes (Stechmann and Cavalier-Smith 2002) and recent analyses of mitochondrial genes (Derelle and Lang 2011). A reinterpretation of the presence/absence of a DHFR-TS gene fusion led to the hypothesis that the root lies between "Unikonta" (Opisthokonta + Amoebozoa) and "Bikonta" (Stechmann and Cavalier-Smith 2003 but see Arisue et al. 2005 and Nozaki et al. 2005). The "Archezoa" root (i.e., between amitochondriate eukaryotes such as Diplomonads and Parabasalids and all remaining eukaryotes) was predicated on the hypothesis that early eukaryotes lacked mitochondria (Cavalier-Smith 1993), a claim that is not supported by the preponderance of evidence (Hirt et al. 1997; Roger et al. 1999; Brinkmann and Philippe 2007). A hypothesis that the root of eukaryotes falls either within or at the base of the Euglenozoa is derived from observations of 12

TABLE 3. Analysis of hypotheses from the literature reveal support for a root between Opisthokonta (or Fungi) and other eukaryotes

	15 w/outgroups		10 w/outgroups		15 no outgroups		10 no outgroups	
Root constraint	Score	Dist.	Score	Dist.	Score	Dist.	Score	Dist.
Opisthokont	3470	—	4440	—	3501	—	5036	—
Fungi	3525	55	4515	75	3488	−13	5042	6
Unikont	3538	68	4533	93	3590	89	5139	103
Plantae	3598	128**	4544	104*	3606	105*	5131	95
Euglenozoa	3618	148*	4564	124	3611	110	5124	88
Amoebozoa	3591	121*	4562	122	3641	140*	5136	100
Excavata	3615	145**	4584	144*	3633	132*	5205	169*
SAR	3561	91	4597	157*	3644	143	5231	195**

Notes: Score = duplication + loss score for most parsimonious tree; Dist = Difference in score from Opisthokont root. Reject hypothesis that alternate roots have equal duplication plus loss scores as tree with the opisthokont rooting, done using Wilcoxon signed-rank test.
*$P < 0.05$ and **$P < 0.01$.

genomic characters that are either absent or reduced in trypanosomes but are widespread in other eukaryotes (Cavalier-Smith 2010). There is also a suggestion that fungi may represent the earliest eukaryotes based on their osmotrophic metabolism and remarkable (at least among eukaryotes) diversity of ATP-producing pathways (Martin et al. 2003).

We assessed alternative hypotheses for the position of the root of a phylogeny by comparing the minimum numbers of gene duplications and losses required under each hypothesis. We analyzed seven alternative hypotheses for the root of eukaryotes. Constraining the tree to a non-opisthokont root resulted in a substantial increase in the number of duplications and losses needed to reconcile the gene trees (Table 3). For example, an additional 68–103 steps (duplications + losses) are needed to generate trees with a root between "unikonts" and all other eukaryotes. The trends are consistent across all analyses in that more than 65 additional duplications or losses are required for all nonopisthokont rootings (Table 3). We also used a Wilcoxon signed-rank test (Wilcoxon 1945) to assess the hypothesis that alternate roots have equal parsimony scores as the best tree rooted on Opisthokonta and find that alternate rootings can be rejected in at least one analysis for all nonopisthokont hypotheses except a "Unikonta" root, which was proposed by Stechmann and Cavalier-Smith (2003).

Synthesis

Rooting the eukaryotic tree of life between opisthokonts and all other eukaryotes challenges the belief that macroscopic eukaryotes are higher forms (i.e., at the crown) than their microbial relatives. Instead, the GTP analyses presented here indicate that there was an early divide among eukaryotes, with one lineage giving rise to animals, fungi, and their microbial relatives whereas the other lineage split into a plethora of intermingled unicellular and multicellular lineages (e.g., plants, red algae, brown algae, slime molds, and water molds). Furthermore, elucidating the root of the tree of eukaryotes is necessary to infer ancestral states and determine directionality in the patterns of eukaryote evolution.

FUNDING

This work was supported by several grants to L.A.K.: NSF Systematics grant (DEB RUI:0919152), NSF Assembling the tree of life grant (DEB 043115), and NIH AREA award (1R15GM081865-01).

ACKNOWLEDGMENTS

This work benefitted from discussions with Daniel Lahr (Smith College/University of Massachusetts), Oliver Eulenstein, and Andre Wehe (Iowa State University) as well as Associate Editor Michael Charleston and anonymous reviewers.

REFERENCES

Adl S.M., Simpson A.G.B., Farmer M.A., Andersen R.A., Anderson O.R., Barta J.R., Bowser S.S., Brugerolle G., Fensome R.A., Fredericq S., James T.Y., Karpov S, Kugrens P., Krug J., Lane C.E., Lewis L.A., Lodge J., Lynn D.H., Mann D.G., McCourt R.M., Mendoza L., Moestrup ∅., Mozley-Standrdge S.E., Nerad T.A., Shearer C.A., Smirnov A.V., Spiegel F.W., Taylor M.F.J.R. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J. Eukaryot. Microbiol. 52:399–451.

Arisue N., Hasegawa M., Hashimoto T. 2005. Root of the eukaryota tree as inferred from combined maximum likelihood analyses of multiple molecular sequence data. Mol. Biol. Evol. 22:409–420.

Baldauf S.L., Roger A.J., Wenk-Siefert I., Doolittle W.F. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science. 290:972–977.

Bansal M.S., Burleigh J.G., Eulenstein O. 2010. Efficient genome-scale phylogenetic analysis under the duplication-loss and deep coalescence cost models. BMC Bioinformatics. 11(Suppl 1):S42.

Brinkmann H., Philippe H. 2007. The diversity of eukaryotes and the root of the eukaryotic tree. Eukaryotic membranes and cytoskeleton: origins and evolution. Berlin (Germany): Springer-Verlag. p. 20–37.

Brown J.R., Doolittle W.F. 1999. Gene descent, duplication, and horizontal transfer in the evolution of glutamyl- and glutaminyl-tRNA synthetases. J. Mol. Evol. 49:485–495.

Burki F., Inagaki Y., Brate J., Archibald J.M., Keeling P.J., Cavalier-Smith T., Sakaguchi M., Hashimoto T., Horak A., Kumar S., Klaveness D., Jakobsen K.S., Pawlowski J., Shalchian-Tabrizi K. 2009. Large-scale phylogenomic analyses reveal that two enigmatic protist lineages, telonemia and centroheliozoa, are related to photosynthetic chromalveolates. Genome Biol. Evol. 1:231–238.

Burleigh J.G., Bansal M.S., Eulenstein O., Hartmann S., Wehe A., Vision T.J. 2011. Genome-scale phylogenetics: inferring the plant tree of life from 18,896 gene trees. Syst. Biol. 60:117–125.

Burleigh J.G., Driskell A.C., Sanderson M.J. 2006. Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. Syst. Biol. 55:426–440.
Cavalier-Smith T. 1993. Kingdom Protozoa and its 18 phyla. Micro. Rev. 57:953–994.
Cavalier-Smith T. 2010. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. Biol. Lett. 6:342–345.
Chaudhary R., Bansal M.S., Wehe A., Fernandez-Baca D., Eulenstein O. 2010. iGTP: a software package for large-scale gene tree parsimony analysis. BMC Bioinformatics. 11:574.
Cotton J.A., Page R.D.M. 2002. Going nuclear: gene family evolution and vertebrate phylogeny reconciled. Proc. R. Soc. Biol. Sci. Ser. B. 269:1555–1561.
Derelle R., Lang F.B. 2011. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. Mol. Biol. Evol. doi: 10.1093/molbev/msr295.
Eddy S.R. 2001. HMMER: profile hidden markov models for biological sequence analysis [Internet]. Ashburn, VA. Available from: http://hmmer.janelia.org/.
Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.
Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 39:783–791.
Gogarten J.P., Kibak H., Dittrich P., Taiz L., Bowman E.J., Bowman B.J., Manolson M.F., Poole R.J., Date T., Oshima T., Konishi J., Denda D., Yoshida M. 1989. Evolution of the vacuolar proton ATPase: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. U.S.A. 86:6661–6665.
Goodman M., Czelusniak J., Moore G.W., Romeroherrera A.E., Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. 28:132–163.
Gorecki P., Tiuryn J. 2007. URec: a system for unrooted reconciliation. Bioinformatics. 23:511–512.
Guigo R., Muchnik I., Smith T.F. 1996. Reconstruction of ancient molecular phylogeny. Mol. Phylogenet. Evol. 6:189–213.
Hampl V., Hug L., Leigh J.W., Dacks J.B., Lang B.F., Simpson A.G.B., Roger A.J. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". Proc. Natl. Acad. Sci. U.S.A. 106:3859–3864.
Hirt R.P., Healy B., Vossbrinck C.R., Canning E.U., Embley T.M. 1997. A mitochondrial Hsp70 orthologue in *Vairimorpha necatrix*: molecular evidence that microsporidia once contained mitochondria. Curr. Biol. 7:995–998.
Iwabe N., Kuma K.I., Hasegawa M., Osawa S., Miyata T. 1989. Evolutionary relationship of archaebacteria, eubacteria and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. U.S.A. 86:9355–9359.
Joly S., Bruneau A. 2009. Measuring branch support in species trees obtained by gene tree parsimony. Syst. Biol. 58:100–113.
Jones D.T., Taylor W.R., Thornton J.M. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275–282.
Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46: 523–536.
Mans B.J., Anantharaman V., Aravind L., Koonin E.V. 2004. Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex. Cell Cycle. 3:1612–1637.
Martin W., Rotte C., Hoffmeister M., Theissen U., Gelius-Dietrich G., Ahr S., Henze K. 2003. Early cell evolution, eukaryotes, anoxia, sulfide, oxygen, fungi first (?), and a tree of genomes revisited. IUBMB Life. 55:193–204.
Nozaki H., Matsuzaki M., Misumi O., Kuroiwa H., Higashiyama T., Kuroiwa T. 2005. Phylogenetic implications of the CAD complex from the primitive red alga Cyanidioschyzon merolae (Cyanidiales, Rhodophyta). J. Phycol. 41:652–657.
Page R.D.M. 2002. Taxonomy, supertrees, and the tree of life. Phylogenetic supertrees: combining information to reveal the tree of life. Dordrecht (The Netherlands): Kluwer Academic Press. p. 247–265.
Parfrey L.W., Barbero E., Lasser E., Dunthorn M., Bhattacharya D., Patterson D.J., Katz L.A. 2006. Evaluating support for the current classification of eukaryotic diversity. PLoS. Genet. 2:e220.
Parfrey L.W., Grant J., Tekle Y.I., Lasek-Nesselquist E., Morrison H.G., Sogin M.L., Patterson D.J., Katz L.A. 2010. Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. Syst. Biol. 59:518–533.
Patterson D.J. 1999. The diversity of eukaryotes. Am. Nat. 154: S96–S124.
Roger A., Morrison H.G., Sogin M.L. 1999. Primary structure and phylogenetic relationships or a malate dehydrogenase gene from *Giardi lamblia*. J. Mol. Evol. 48:750–755.
Roger A.J., Hug L.A. 2006. The origin and diversification of eukaryotes: problems with molecular phylogenetics and molecular clock estimation. Philos. Trans. R. Soc. B Biol. Sci. 361: 1039–1054.
Sanderson M.J., McMahon M.M. 2007. Inferring angiosperm phylogeny from EST data with widespread gene duplication. BMC Evol. Biol. 7(Suppl 1):S3.
Sogin M.L., Gunderson J.H., Eldwood H.J., Alonso R.A., Peattie D.A. 1989. Phylogenetic meaning of the Kingdom concept: an unusual ribosomal RNA from the *Giardia lamblia*. Science. 243: 75–77.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.
Stechmann A., Cavalier-Smith T. 2002. Rooting the eukaryote tree by using a derived gene fusion. Science. 297:89–91.
Stechmann A., Cavalier-Smith T. 2003. Phylogenetic analysis of eukaryotes using heat-shock protein Hsp90. J. Mol. Evol. 57:408–419.
Tekle Y.I., Parfrey L.W., Katz L.A. 2009. Molecular data are transforming hypotheses on the origin and diversification of eukaryotes. Bioscience. 59:471–481.
Templeton A.R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution. 37:221–244.
Van de Peer Y., De Wachter R. 1997. Evolutionary relationships among the eukaryotic crown taxa taking into account site-to-site variation in 18S rRNA. J. Mol. Evol. 45:619–630.
Wehe A., Bansal M.S., Burleigh J.G., Eulenstein O. 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics. 24:1540–1541.
Wilcoxon F. 1945. Individual comparisons by ranking methods. Biometrics Bull. 1:80–83.
Yoon H.S., Grant J., Tekle Y.I., Wu M., Chaon B.C., Cole J.C., Logsdon J.M., Patterson D.J., Bhattacharya D., Katz L.A. 2008. Broadly sampled multigene trees of eukaryotes. BMC Evol. Biol. 8:14.